

Locally Stationary Distributions: A Framework for Analyzing Slow-Mixing Markov Chains*

Kuikui Liu EECS MIT Cambridge, USA liukui@mit.edu	Sidhanth Mohanty EECS MIT Cambridge, USA sidm@mit.edu	Prasad Raghavendra EECS UC Berkeley Berkeley, USA raghavendra@berkeley.edu	Amit Rajaraman EECS MIT Cambridge, USA amit_r@mit.edu	David X. Wu EECS UC Berkeley Berkeley, USA david_wu@berkeley.edu
---	---	--	---	--

Abstract—Many natural Markov chains fail to mix to their stationary distribution in polynomially many steps. Often, this slow mixing is inevitable since it is computationally intractable to sample from their stationary measure.

Nevertheless, Markov chains can be shown to always converge quickly to measures that are *locally stationary*, i.e., measures that don't change over a small number of steps. These locally stationary measures are analogous to local minima in continuous optimization, while stationary measures correspond to global minima.

While locally stationary measures can be statistically far from stationary measures, do they enjoy provable theoretical guarantees that have algorithmic implications? We study this question in this work and demonstrate three algorithmic applications of locally stationary measures:

- 1) We show that Glauber dynamics on the hardcore model can be used to find large independent sets in triangle-free graphs of bounded degree.
- 2) We prove that Glauber dynamics on the Ising model defined by a spiked matrix model finds a vector with constant correlation with the planted spike.
- 3) We show that for sufficiently large constant signal-to-noise ratio, Glauber dynamics on the Ising model finds a vector that has constant correlation with the hidden community vector.

In other words, Glauber dynamics subsumes the spectral method for *spiked Wigner* and *community detection*, by weakly recovering the planted spike.

The full version of this paper can be found on arXiv (arXiv ID: 2405.20849).

Index Terms—sampling, inference, spiked Wigner, stochastic block model, Ising model, statistical physics

I. INTRODUCTION

Markov chains are a fundamental algorithmic primitive that are widely applied towards sampling and counting tasks. There is a rich body of literature devoted to understanding worst-case *mixing times* of Markov chains, i.e., the number of steps required for the distribution of the chain to approach its stationary measure started from an arbitrary initialization. For some highlights in this area, see, e.g., the contents and references in [1]–[5].

A.R. is grateful for the support of an Akamai Presidential Fellowship. D.X.W. is grateful for the support of NSF Graduate Research Fellowship DGE-146752.

Unfortunately, many natural Markov chains fail to mix rapidly from worst-case initializations, in that it takes a super-polynomial number of steps to reach stationarity. Structurally, this is due to the presence of cuts in the state space with very small conductance. Often, slow mixing is inevitable since sampling from the stationary measure is known to be computationally hard (say NP-hard); see, e.g., [6]–[8]. Nevertheless, it has been empirically observed that certain simple and local Markov chains like *Glauber dynamics* succeed at optimization and inference tasks even when they are not known to mix, such as finding satisfying assignments to SAT formulas [9], [10], and clustering stochastic block models [11], [12]. This suggests that local Markov chains like Glauber dynamics can have algorithmic applications, even if they *fail* to mix rapidly, which raises our main line of inquiry.

Question 1.1. What is the long-term behavior of Markov chains that do not mix rapidly? How can slow-mixing Markov chains be harnessed for optimization and inference?

To lay out the motivation, it is useful to draw an analogy to continuous optimization. Gradient descent, the canonical algorithm in optimization, converges efficiently to a global minimum if the objective function and parameter space is convex. However, non-convex objective functions and parameter spaces come up often both in theory and practice, and finding global minima can even be provably intractable. On the other hand, gradient descent can be shown to always converge quickly to a local minimum, or more precisely, a first-order stationary point. Moreover, these local minima are useful in practice, and can admit non-trivial theoretical guarantees. See [13] for a comprehensive coverage of analyzing gradient-based optimization methods, and see, e.g., [14]–[17] and the references therein for non-trivial theoretical guarantees on local minima of gradient descent in the context of machine learning.

Analogously, in the context of sampling, certain random walks can be shown to mix rapidly to their stationary measures. However, other random walks can also be shown to mix slowly from worst-case initializations; these are akin to the “hard” nonconvex optimization problems. Despite this, one can show that *any* Markov chain satisfying fairly generic conditions converges to analogs of local minima that we term

locally stationary distributions (see Definition I.2). Intuitively, a locally stationary distribution corresponds to the stationary measure conditioned on a subset of states that are sparsely connected to the rest of the state space under the Markov chain.

This raises the question of whether these locally stationary measures obey theoretical guarantees that are useful for solving problems in optimization or inference.

A. Locally stationary distributions

Let P be the transition matrix of a time-reversible Markov chain on a state space Ω , where $P[x, y]$ denotes the transition probability from x to y . Let π be a stationary distribution w.r.t. P .

Analogous to local minima in optimization, a locally stationary measure ν is one started at which the Markov chain P remains nearly stationary, i.e., makes little progress. We will use the *KL divergence* to the stationary distribution, denoted $\text{KL}(\nu\|\pi)$, as a measure of progress. This leads to the following definition of an ε -locally stationary measure.

Definition I.2. A probability measure ν on Ω with density f relative to π is said to be ε -locally stationary with respect to P if

$$\mathcal{E}(f, \log f) := \sum_{x, y \in \Omega} P[x, y] \cdot (f(x) - f(y)) \cdot \log \frac{f(x)}{f(y)} \leq \varepsilon$$

The *Dirichlet form* $\mathcal{E}(f, \log f)$ measures the rate at which the Markov chain progresses towards the stationary distribution. In particular, for a continuous-time version of the Markov chain P , if ν_t denotes the measure at time t , we have the following well-known fact [18]:

$$\frac{d}{dt} \text{KL}(\nu_t\|\pi) = -\mathcal{E}(f_t, \log f_t).$$

As an immediate consequence, we observe that the Markov chain is typically on ε -locally stationary measures over time. Formally, we have the following claim.

Theorem I.3. Fix a time-reversible Markov chain P with a stationary measure π , any starting distribution ν_0 , and $\varepsilon, \delta > 0$. Let $T = \frac{1}{\delta\varepsilon} \cdot \log\left(\frac{1}{\pi_{\min}}\right)$. Then, for a time $t \sim [0, T]$ chosen uniformly at random, the distribution ν_t at time t is ε -locally stationary with respect to P with probability at least $1 - \delta$.

The main conceptual contribution of our work is the following meta-principle for showing that locally stationary distributions solve optimization and inference problems.

Prove that sampling from the true stationary distribution solves the optimization or inference problem of interest, and additionally, does so for “local” reasons.

This principle is best illustrated by discussing our algorithmic applications of locally stationary distributions.

Independent sets in triangle-free graphs. It is easy to see that any graph G on n vertices with degree bounded by d has an independent set of size $\frac{n}{d+1}$, a bound which is tight

for the union of $(d+1)$ -sized cliques. Ajtai, Komlós, and Szemerédi [19] showed that when G is triangle-free, the size of the maximum independent set guaranteed to exist increases to $\Omega\left(n \cdot \frac{\log d}{d}\right)$. Shearer [20] gave an alternate proof which pins down the leading constant to $1 - o_d(1)$ and relaxes the assumption of bounded maximum degree to bounded average degree.

It is also well-known that a uniformly random independent set in a triangle-free graph of maximum degree d has expected size at least $\Omega\left(n \cdot \frac{\log d}{d}\right)$; see, e.g., [21, Proposition 1, Page 272]. Hence, it is natural to wonder whether *Glauber dynamics* with respect to the uniform measure over independent sets finds such a large independent set. From a given independent set $I \subseteq V$, the transitions of Glauber dynamics can be described as follows:

- 1) Sample a uniformly random vertex $v \in V$.
- 2) If $I \cup \{v\}$ is an independent set, then go to $I' = I \cup \{v\}$ with probability $1/2$ and $I' = I \setminus \{v\}$ with probability $1/2$.
- 3) If $I \cup \{v\}$ is not an independent set, go to $I' = I \setminus \{v\}$ with probability 1 .

Notably, this Markov chain requires $\exp(\Omega(n))$ steps to mix [22] as soon as $d \geq 6$. In fact, the problem of sampling a uniformly random independent set on a graph of maximum degree d becomes NP-hard [6], [7] in this regime, even if triangle-freeness is assumed [23].

Despite these hardness results for the corresponding sampling problems, we show that the above Markov chain can be used to find independent sets of size $\Omega\left(n \cdot \frac{\log d}{d}\right)$ in triangle-free graphs of maximum degree bounded by d . Specifically, we show the following result.

Theorem I.4. Let G be a triangle-free graph on n vertices with maximum degree bounded by d . Let I be an independent set in G that arises from Glauber dynamics run for $O(nd^4)$ time. Then the expected size of I is at least $\frac{1-o_d(1)}{4} \cdot n \cdot \frac{\log d}{d}$.

Remark I.5. In fact, one can prove that Glauber dynamics at “fugacity” $\frac{1}{\log d}$ finds an independent set of size $(1 - o_d(1)) \cdot n \cdot \frac{\log d}{d}$ by combining our proof method with that of [24].

As mentioned before, we know that the expected size of a uniformly random independent set satisfies the above lower bound. However, the Glauber dynamics chain does not mix rapidly, and hence does not produce samples from the truly uniform distribution. Instead, it samples from a locally stationary distribution with respect to the Markov chain. Our key insight is that the same proof also goes through for an independent set sampled from a locally stationary distribution with respect to Glauber dynamics.

To give a sense of how local stationarity is used, we briefly discuss the proof. The proof from [21] that the expected size of a uniformly random independent set $\Omega\left(n \cdot \frac{\log d}{d}\right)$ argues that for any vertex v , and for any pinning $x_{v \cup N(v)}$ of the independent set outside v and its neighbors, either:

- the uniform distribution conditioned on the pinning chooses v with probability $\gtrsim \frac{\log d}{d}$, or
- it chooses $\gtrsim \log d$ neighbors of v in expectation.

Thus, each vertex can be charged $\Omega\left(\frac{\log d}{d}\right)$ vertices on average in the independent set. Observe that the above sketch of the argument goes through even if the distribution is not truly uniform but merely has conditional marginals matching the uniform distribution, which is a property we can show holds for locally stationary distributions (Lemma III.5).

Using similar arguments, one can establish that given a triangle-free graph with maximum degree d , Glauber dynamics run for $\text{poly}(n)$ many steps on the antiferromagnetic Ising model on G with inverse temperature $\frac{1}{\sqrt{d}}$ recovers a cut of relative size $\frac{1}{2} + \Omega\left(\frac{1}{\sqrt{d}}\right)$.

Weak recovery in spiked models. Beyond independent sets, we also study the performance of Glauber dynamics for statistical inference tasks. Consider the central class of Bayesian models for principal component analysis (PCA) known as *spiked random matrix models*, which consist of a matrix $M \in \mathbb{R}^{n \times n}$ given by

$$M = \begin{array}{c} \lambda \\ \uparrow \\ \text{signal strength} \end{array} \cdot \begin{array}{c} vv^\top \\ \uparrow \\ \text{signal} \end{array} + \begin{array}{c} W \\ \uparrow \\ \text{noise} \end{array}.$$

The general algorithmic question is to approximately recover the signal (a unit vector v) under appropriate assumptions about the noise (W) and the signal strength (λ). More precisely:

Problem I.6 (Weak recovery in spiked matrix models). For a unit norm signal vector $v \in \mathbb{R}^n$, signal strength $\lambda \in \mathbb{R}$, and noise matrix $W \in \mathbb{R}^{n \times n}$, given $M = \lambda \cdot vv^\top + W$, give an efficient algorithm to extract a unit norm estimate \hat{v} such that $\langle v, \hat{v} \rangle \geq \Omega(1)$.

In the situation where W is a Wigner matrix, this model, known as the *spiked Wigner model*, has been a subject of extensive study. The work of [25] determined that once $\lambda > 1$, a spectral algorithm based on computing the top eigenvector succeeds at weak recovery. There is a fairly large body of work on the spiked Wigner model, towards characterizing optimal estimation error, efficient algorithms and its generalizations to rank larger than one [26]–[33]. When the prior distribution over v is the uniform distribution over $\left\{\pm \frac{1}{\sqrt{n}}\right\}^n$ and \mathbb{S}^{n-1} , there are efficient algorithms that even achieve the maximum information-theoretically achievable correlation $|\langle v, \hat{v} \rangle|$, based on approximate message passing [34], [35], and algorithmic stochastic localization [36].

In the case where the prior distribution is on the hypercube, given the matrix M , this posterior is described by an Ising model — a probability distribution $\mu_{\beta M}$ over $\{\pm 1\}^n$ defined by the following proportionality relation for a suitably chosen $\beta > 1$:

$$\mu_{\beta M}(x) \propto \exp\left(\frac{1}{2}\langle x, \beta Mx \rangle\right) \text{ for all } x \in \{\pm 1\}^n.$$

Sampling from the above posterior distribution is desirable as it achieves the maximum information-theoretically achievable correlation. The canonical algorithm for sampling is to run the Glauber dynamics Markov chain, but unfortunately, provable guarantees for Glauber dynamics are currently lacking. Thus, a natural question en route is: *does Glauber dynamics for the Ising model $\mu_{\beta M}$ weakly recover the signal in polynomial-time?*

We make progress towards answering this question affirmatively in this work by showing that Glauber dynamics at a slightly higher temperature than the posterior distribution succeeds for a broad family of settings.

Formally, we show the following result:

Theorem I.7. *Let W be a matrix with $\kappa \leq W \leq 1 - \kappa$, and v a unit vector with $\|v\|_\infty \leq L/\sqrt{n}$ for some positive constant L . Let P denote the kernel of the Glauber dynamics chain with stationary distribution $\mu_{W+\lambda vv^\top}$, and x_0 an arbitrary point on $\{\pm 1\}^n$. There exists a large enough constant $\lambda > 0$ such that for $T = \tilde{\Theta}(n^4)$, and for $t \sim [0, T]$, with probability $1 - o(1)$ we have:*

$$\mathbf{E}_{\mathbf{x} \sim P^t \delta_{x_0}} [|\langle \mathbf{x}, v \rangle|] \geq \left(\frac{1}{L} \cdot \kappa \exp\left(-\frac{1}{\kappa}\right) - o(1) \right) \cdot \sqrt{n}.$$

A natural approach to recover the signal v from the matrix M is the spectral method, which amounts to computing (even approximately) the eigenvector corresponding to the largest eigenvalue for the matrix M . At a high-level, the above theorem demonstrates that Glauber dynamics can *simulate* the spectral method in certain regimes. We further expect Glauber to achieve weak recovery when run for $T = n^{1+o(1)}$ steps, but we leave this open as a direction for future improvement.

Remark I.8. The above model of choice captures several commonly considered models of study in the algorithms and complexity of statistical inference, such as the *spiked Wigner model* [25], and *random/planted 2XOR* (see, e.g., [37] and the references within).

Remark I.9. In the Rademacher spiked Wigner model, where $W \sim \text{GOE}(n)$ and $v \sim \left\{\pm \frac{1}{\sqrt{n}}\right\}^n$, the posterior has the form

$$\Pr[v|M] \propto \exp\left(-\frac{n}{2} \|M - \lambda vv^\top\|_F^2\right) \propto \exp\left(\frac{\lambda n}{2} \cdot v^\top M v\right).$$

The above is an Ising model, and suggests Glauber dynamics as a natural algorithm for weak recover. We note that the Ising model we shall analyze will be a higher temperature version of the above, that is, a distribution with density proportional to $\exp\left(\frac{\beta n}{2} \cdot v^\top M v\right)$ for some $\beta < \lambda$ (as opposed to the “correct” value λ). Interestingly, such recovery guarantees were not previously known, even if one *information theoretically* samples from the higher temperature Ising model.

Stochastic block model. Another case of interest is one where the Ising model M arises from a stochastic block model. To describe this result, we first begin by defining the two-community stochastic block model.

Definition I.10 (2-community stochastic block model). Let $d, \lambda \in \mathbb{R}$ be fixed parameters such that $\lambda^2 \leq d$. The distribution $\text{SBM}(n, d, \lambda)$ is defined over pairs $(\sigma, G) \in \{\pm 1\}^n \times \{0, 1\}^{n \times n}$ generated as follows.

Let $\sigma \in \{\pm 1\}^n$ be a *signal* vector drawn uniformly at random (i.e. the prior is uniform). Given σ , we draw a random graph G by including an edge between $u, v \in [n]$ independently with probability $\frac{d+\lambda\sqrt{d}}{n}$ if $\sigma(u) = \sigma(v)$, and with probability $\frac{d-\lambda\sqrt{d}}{n}$ otherwise.

In a general stochastic block model, the signal vector σ can be over a larger finite alphabet $[q]$, and the probability of including an edge between $u, v \in [n]$ is an arbitrary function of $\sigma(u), \sigma(v)$. In this work, we will use the term stochastic block model (SBM) to refer exclusively to the special case of two communities as defined above.

Remark I.11. The 2-community stochastic block model can be viewed as a special case of a spiked matrix model where M is a highly sparse matrix. Due to the sparsity of M , this spiked matrix model falls outside the scope of Theorem I.7, as the “noise” part fails to satisfy the spectral bound.

The weak recovery problem for stochastic block model is that of recovering a labelling $\hat{\sigma}$ given the graph G such that $\hat{\sigma}$ has non-trivial correlation with the true signal σ . More precisely, an algorithm for weak recovery is required to find a $\hat{\sigma}$ such that $\frac{1}{n}|\langle \hat{\sigma}, \sigma \rangle| \geq \Omega(1)$.

Starting with the work of Decelle, Krzakala, Moore, & Zdeborova [38] that posited broad conjectures about these models, an extensive body of work has emerged over the past decade. For the case of 2 communities, [38] posited that weak-recovery is possible if and only if the signal strength $\lambda^2 > 1$. This coincides with the *Kesten–Stigum threshold*, a threshold for broadcast processes on trees studied in the works of Kesten and Stigum [39], [40]. The works of Mossel, Neeman, & Sly [41] and Massoulié [42] confirmed the algorithmic side, namely that weak recovery can be solved efficiently above the KS threshold with a spectral algorithm, while [43] showed impossibility below the threshold. We refer the reader to the survey of Abbe [44] for a detailed treatment of the literature on community detection.

We show that Glauber dynamics succeeds at weak recovery when the signal strength is a constant factor above the Kesten–Stigum threshold.

Theorem I.12. *There exist constants $\lambda_0, \beta, c > 0$ such that for all λ satisfying $|\lambda| \geq \lambda_0$, for $(\sigma, G) \sim \text{SBM}(n, d, \lambda)$, with probability $1 - o(1)$ over the randomness of (σ, G) , the following holds.*

Let P denote the kernel of Glauber dynamics with stationary distribution $\mu_{\frac{\beta}{\sqrt{d}}(A_G - \frac{d}{n}\mathbf{1}\mathbf{1}^\top)}$ and x_0 an arbitrary point on $\{\pm 1\}^n$. For $T = \tilde{\Theta}(n^{4+o(1)})$, and for $t \sim [0, T]$, with probability $1 - o(1)$, we have:

$$\mathbf{E}_{\mathbf{x} \sim P^t \delta_{x_0}}[|\langle \mathbf{x}, \sigma \rangle|] > cn.$$

Remark I.13. Similar to Remark I.9, the posterior distribution $\sigma|G$ to solve the recovery problem in the stochastic block model is an Ising model (see, e.g., [45, Eq. (7)]):

$$\Pr[\sigma|G] \propto \prod_{ij \in E(G)} \left(\frac{d + \lambda\sqrt{d}}{d - \lambda\sqrt{d}} \right)^{(1+\sigma_i\sigma_j)/2} \cdot \prod_{ij \notin E(G)} \left(\frac{1 - \frac{d + \lambda\sqrt{d}}{n}}{1 - \frac{d - \lambda\sqrt{d}}{n}} \right)^{(1+\sigma_i\sigma_j)/2}$$

For large d , the Ising model that we analyze is approximately equal to a higher temperature version of the above true posterior.

Although spectral algorithms for weak recovery were already known in all the cases listed above, understanding the power of Glauber dynamics is interesting in its own right. It is arguably a more natural algorithm than spectral methods in the context of a Bayesian estimation problem like stochastic block models. In particular, Glauber dynamics remains locally consistent with the underlying probabilistic model at every vertex. On the other hand, a spectral algorithm that computes the top eigenvector maximizes a global objective, while crudely approximating the local features of the probabilistic model.

Finally, our analysis for spiked models establishes a direct correspondence between locally stationary measures for Glauber dynamics and fixed points of a Markov chain over the one-dimensional real line \mathbb{R} (related to the *restricted Gaussian dynamics* Markov chain). This correspondence may pave the way for a much tighter analysis to establish that Glauber dynamics achieves information theoretically optimal recovery in some of these models. To elucidate further on this correspondence, we will give a brief technical overview here.

B. Technical overview

In Section III, we derive a few basic properties of locally-stationary measures. This is followed by the result on independent sets presented in Section IV as a warmup. We omit the proofs of Theorems I.7 and I.12, and refer the reader to the full version of the paper for details.

In this technical overview, we will focus on the inference problem in spiked matrix models. For a matrix $M \in \mathbb{R}^{n \times n}$ and a vector $h \in \mathbb{R}^n$, we will use $\mu_{M,h}$ to denote the distribution over $\{\pm 1\}^n$ defined as

$$\mu_{M,h}(x) \propto \exp\left(\frac{1}{2}\langle x, Mx \rangle + \langle h, x \rangle\right),$$

and μ_M to denote the distribution $\mu_{M,0}$.

Consider the stationary measure μ_M for a spiked matrix $M = \lambda vv^\top + W$. We outline the proofs of Theorems I.7 and I.12 here, which consist of two parts.

- First, we show that locally stationary distributions with respect to Glauber dynamics over $\{\pm 1\}^n$ are also locally stationary with respect to the *restricted Gaussian dynamics* Markov chain.
- Next, we show that samples from locally stationary distributions for RGD achieve weak recovery.

Let us first recall the definitions of Glauber dynamics and Restricted Gaussian dynamics for Ising models.

Definition I.14 (Glauber dynamics). Glauber dynamics with respect to a distribution π over $\{\pm 1\}^n$ is a Markov chain on $\{\pm 1\}^n$, where a transition from x is given by the following:

- Sample index i uniformly from $[n]$.
- Transition to $x^{\oplus i}$ with probability $\frac{\pi(x^{\oplus i})}{\pi(x) + \pi(x^{\oplus i})}$, and stay at x otherwise. Here, $x^{\oplus i}$ denotes x with the i th bit flipped.

Definition I.15 (Restricted Gaussian dynamics; cf. [46]–[48]). Consider the joint random variable (\mathbf{x}, \mathbf{z}) where $\mathbf{x} \sim \mu_M$, and $\mathbf{z}|\mathbf{x} := (\lambda \langle v, \mathbf{x} \rangle + \sqrt{\lambda} \mathbf{g}) \cdot v$ for $\mathbf{g} \sim \mathcal{N}(0, 1)$. *Restricted Gaussian dynamics* (RGD) is a Markov chain on $\{\pm 1\}^n$ where for any x , the transition to x' is sampled as follows:

- Sample $\mathbf{z}|x$.
- Sample $\mathbf{x}'|\mathbf{z}$.

Remark I.16. By definition, the above Markov chains are ergodic and reversible with respect to π and μ_M respectively, and so asymptotically converge to them as their stationary distributions.

Remark I.17. We should think of $\mathbf{z}|x$ as being a noisy surrogate for how well \mathbf{x} correlates with the hidden direction v .

Informally, we prove the following correspondence between locally stationary distributions for Glauber dynamics and locally stationary distributions for RGD. In fact, this correspondence is a consequence of a more generic statement; refer to Lemma III.8 for details.

Lemma I.18 (Informal version of Lemma III.8). *Let ν be a distribution over $\{\pm 1\}^n$ that is ε -locally stationary under Glauber dynamics for μ_M . Suppose for every $z \in \mathbb{R}$, Glauber dynamics for the distribution of $\mathbf{x}|z$ is “well-expanding”, and $\log \frac{1}{\mu_M(x)} \leq \text{poly}(n)$ for all $x \in \{\pm 1\}^n$, then ν is $\varepsilon \cdot \text{poly}(n)$ -locally stationary under restricted Gaussian dynamics.*

To conclude that ν is locally stationary under restricted Gaussian dynamics, it suffices to verify the structural properties of μ_M . The lower bound on the minimum probability follows from upper and lower bounds on the values that the Hamiltonian can achieve. To show that Glauber dynamics for $\mathbf{x}|z$ is well-expanding, we must investigate the structure of this distribution further. A simple calculation reveals that the distribution of $\mathbf{x}|z$ is, in fact, the Ising model $\mu_{W,z}$. In the setting of Theorem I.7, where the spectral diameter of W is bounded by 1, prior works [48]–[51] prove that $\mu_{W,z}$ always satisfies a “modified log-Sobolev inequality” (our relevant notion of “well-expanding”). In the setting of Theorem I.12, where W is a centered stochastic block model, a similar result is proved in a companion paper [52].

Remark I.19. This decomposition of μ_M into a mixture of other Ising models is well-known in the literature by the name *Hubbard–Stratonovich transformation* [53]; see also [54].

In summary, we showed if ν is locally stationary with respect to Glauber dynamics, then it is also locally stationary for the RGD chain

$$\mathbf{x} \rightarrow \mathbf{z}|\mathbf{x} \rightarrow \mathbf{x}'|\mathbf{z}.$$

Thus, it suffices to prove that $\mathbf{E}_{\mathbf{x} \sim \nu} |\langle \mathbf{x}, v \rangle|$ is bounded away from 0 for any distribution ν that is locally stationary for RGD. The two ingredients that go into proving this are:

- A generic principle that says: if ν is locally stationary for a Markov chain P , then for any bounded function f , $|\mathbf{E}_{\mathbf{x} \sim \nu} f(\mathbf{x}) - \mathbf{E}_{\mathbf{x} \sim P\nu} f(\mathbf{x})|$ is small (Corollary III.3).
- If the correlation of a distribution ν is too close to 0, then a single step of RGD causes a significant boost in correlation, which means ν cannot be locally stationary. In particular, for $f(x) = |\langle x, v \rangle|$, if $\mathbf{E}_{\mathbf{x} \sim \nu} f(\mathbf{x})$ is too close to 0, then $\mathbf{E}_{\mathbf{x} \sim P\nu} f(\mathbf{x}) - \mathbf{E}_{\mathbf{x} \sim \nu} f(\mathbf{x})$ is nontrivially large, which means any locally stationary distribution ν must achieve large correlation.

See Section 5.2 of the full version of the paper for the details of this argument.

C. Related work

Motivated by statistical physics, the phenomenon of *metastability* of random walks has been extensively studied. We refer the reader to the monograph by Bovier & Hollander [55] for related literature. The notion of metastability in [55] appears to be a slightly stricter notion than local stability, and thus does not generically hold for all reversible Markov chains.

In the context of sampling distributions over a continuous domain, Balasubramanian, Chewi, Erdogdu, Salim & Zhang [56] showed that the Langevin Monte Carlo algorithm always outputs a sample from a distribution whose relative Fisher information is small. This is the continuous sampling analog of convergence of gradient descent to approximate first-order stationary points. Building on these ideas, Cheng, Wang, Zhang & Zhu [57] study the notion of *conditional mixing* for Langevin and Glauber dynamics and apply it to efficiently sample from Gaussian mixtures.

Our analysis of Glauber dynamics borrows ideas from a recent line of works on sampling from Ising models. Glauber dynamics for an Ising model defined by a matrix M was shown to mix quickly if eigenvalues of M lie within an interval of length 1 [50], [51]. This is sharp, as evidenced by the Curie–Weiss model $M = \frac{\beta}{n} \mathbf{1} \mathbf{1}^\top$. Stronger evidence for hardness of sampling beyond this spectral criterion was recently provided by Kunisky [58], based on a reduction to a certain statistical hypothesis testing problem. Koehler, Lee & Risteski [54] devised more sophisticated algorithms based on simulated tempering and variational inference to sample from Ising models when they have constantly many eigenvalues outside an interval of length 1.

Besides the question of fast mixing and metastability, the problem of how well MCMC-based algorithms perform for optimization and inference tasks was recently studied in several works. Chen, Mossel & Zadik [59] proved that when initialized at the empty set, natural Metropolis chains on cliques fail

to find cliques of sublinear size in polynomial time, even if such a clique is planted inside the Erdős–Rényi random graph $G(n, 1/2)$. This is despite there being an abundance of algorithms which can recover a planted clique of size down to $O(\sqrt{n})$. Nevertheless, MCMC-based algorithms were redeemed in a more recent work of Gheissari, Jagannath & Xu [60] using a more carefully designed chain and initialization. In a recent work, Sellke [61] proved that low-temperature Langevin dynamics achieves the conjectured computational threshold for optimizing pure spherical spin glass models.

D. Open problems

We conclude with several open directions, which we believe may be amenable to the framework of locally stationary distributions.

Bayesian inference via MCMC. First, there is the direction of pushing our results further in the setting of SBM. To set the scene, let $\pi(x) \propto \exp(H(x))$ be the true posterior for SBM, where $H(x)$ is the SBM Hamiltonian (see Remark I.13 for an explicit formula). It is well known that optimal recovery is achieved information theoretically by sampling from π (see, e.g., [45, Section 4]). However, it takes exponential time to mix to π from a worst-case initialization. On the other hand, in Theorem I.12, we achieve weak recovery by running Glauber dynamics on the density $\pi_\beta(x) \propto \exp(\beta H(x))$ for some (constant) β strictly smaller than 1. It is natural to investigate whether a sampling algorithm based on *simulated annealing*, i.e. running Glauber dynamics by varying the temperature over time, can succeed at sampling from π . Our main result can be viewed as a modest step in this direction, as we show that running the chain for $\text{poly}(n)$ steps at a mismatched temperature already gives a warm start for correlation.

Problem I.20 (Optimal recovery for stochastic block model). Can an instance of simulated annealing sample from π ?

Computationally optimal inference. The k -community stochastic block model is known to undergo an *information-computation gap* when $k \geq 5$ (see, e.g., [21]). Specifically, for every $k \geq 5$, there exists a choice of degree d and signal-to-noise ratio λ for which weak recovery is information-theoretically possible, but likely impossible for efficient algorithms [62]. This gap admits the construction of SBM instances where weak recovery is tractable, but information-theoretically optimal recovery is intractable to efficient algorithms.

Example I.21. Consider a 10-community block model obtained by taking two disjoint 5-community block model graphs, and planting a sparse bisection between them. The planted bisection should be sparse enough so it is clearly detectable to efficient algorithms. However, the parameters d and λ for the 5-community models are chosen to be in the intractable regime. An information-theoretically optimal algorithm achieves weak recovery within each 5-community model. At the same time, algorithms for the 2-community block model achieve weak recovery in the 10-community

model, since they can find the planted bisection, and correctly classify vertices as belonging to either $\{1, 2, 3, 4, 5\}$ or $\{6, 7, 8, 9, 10\}$.

In settings such as the above, information-theoretically optimal inference is hard, but weak recovery is still tractable. This motivates the study of *computationally optimal inference algorithms* — algorithms that achieve the best guarantees possible in polynomial time. For such problems, the Glauber dynamics chain must necessarily fail to mix rapidly to the posterior distribution, but perhaps the locally stationary distribution it samples from can achieve the computationally optimal recovery guarantees?

Problem I.22 (Computationally optimal inference). Is (annealed) Glauber dynamics a computationally optimal algorithm for the k -community SBM and, more generally, for random CSPs with planted solutions?

Metastable states. Local stationarity is a generic property of any time-reversible Markov chain, so a priori there is no reason to expect that a locally stationary distribution ν has any nice properties. For example, if we run the Markov chain for T steps, Theorem I.3 guarantees an ε -locally stationary distribution ν where $\varepsilon = O(1/T)$, and the simple random walk on the n -vertex cycle graph demonstrates that this is tight if $T = o(n^2)$. This suggests the following natural questions. Suppose the stationary distribution π is a Gibbs distribution on $\{\pm 1\}^n$. Under what additional structural assumptions on π can we both obtain $\varepsilon = o(1/T)$ for sufficiently large $T = \text{poly}(n)$ and endow ν with a physical or geometric interpretation?

The notion of metastable states for Gibbs distributions in statistical physics [55] suggests a conceptual path forward towards this goal. In particular, one might hope to show that a locally stationary distribution is close to a metastable state, i.e. a conditional Gibbs distribution restricted to a metastable set of configurations.

Problem I.23 (Metastability). Suppose π is a Gibbs distribution which has a metastable subset S with exponentially small conductance. Let ν be the locally-stationary distribution after running Glauber (or Langevin) dynamics for $\text{poly}(n)$ steps with uniform initialization in S . Is ν close to the conditional Gibbs distribution π_S , e.g., $\text{KL}(\nu \parallel \pi_S) = o(1)$?

For a concrete setting, suppose π is a spherical spin glass in the *shattering* regime [63]. Can one show that in $\text{poly}(n)$ time, Langevin dynamics with uniform initialization remains stuck in the clusters identified there?

Cavity method. The *cavity method*, and the related *replica method*, originated in physics to predict the properties of various Gibbs distributions. Some striking achievements of this heuristic in producing accurate predictions are the Parisi formula [64], [65], and the k -SAT satisfiability threshold [66]–[68]. It was also employed in the work of Decelle, Kražákala, Moore, & Zdeborová [38] to conjecture the Kesten–Stigum threshold as the computational threshold for SBM.

Of particular interest to us are the works of Coja-Oghlan, Krzakala, Perkins & Zdeborová [69]–[71], that characterize the recovery rate that the optimal estimator, namely sampling from the Gibbs distribution, achieves for various planted constraint satisfaction problems. Their proofs use fairly minimal properties of the Gibbs distribution. More concretely, for a graph G and an assignment x , let $H_G(x)$ be a Hamiltonian, and let π_G be the corresponding Gibbs distribution. Their proofs rely on the following properties satisfied by Gibbs distributions.

- (Gibbs ratios) For any graph G and any vertex v :

$$\frac{\pi_G(x)}{\pi_{G \setminus v}(x)} \propto \exp(H_G(x) - H_{G \setminus v}(x)),$$

- (Approximate independence) For random G and v , the marginals of the neighbors of v are approximately independent in $\pi_{G \setminus v}$.

If one can show that a family of locally stationary distributions also satisfy the Gibbs ratios up to a multiplicative error and approximate independence on a random graph G , then one could hope to port over the cavity method predictions and their rigorous proofs in a black-box fashion.

Problem I.24 (Cavity method). Let $\{\nu_G\}_{G \in \text{graphs}}$ be a family of locally stationary distributions where ν_G arises from running the Glauber dynamics for π_G for time- T initialized at the uniform distribution. For a random graph G and random vertex v , do ν_G and $\nu_{G \setminus v}$ satisfy, up to small error, the Gibbs ratios and approximate independence properties?

Beyond average-case models. Our work proves that Glauber dynamics recovers planted spikes when the input matrix has a clean “signal + noise” structure. Recently, there has been a flurry of work on inference in *semirandom models*; see, e.g., [72], [73] and the references within, where it is possible to extract the hidden signal using semidefinite programming-based algorithms. A natural direction is to investigate whether Glauber dynamics solves semirandom inference problems.

Problem I.25 (Semirandom models). Does Glauber dynamics succeed at finding solutions to semirandom planted CSPs, or large cliques in semirandom graphs with planted cliques as is done in the works of [73] and [72] respectively?

In a similar vein, semidefinite programming has been phenomenally successful at solving dense CSPs, and more generally CSPs on graphs with low threshold-rank [74]. A reason to believe that local algorithms perform well is Theorem I.7, where we show that Glauber dynamics can recover rank-1 spikes in threshold-rank-1 matrices.

Problem I.26 (CSPs on low threshold-rank graphs). Does running Glauber dynamics give a PTAS for Max Cut on a dense graph?

II. PRELIMINARIES

We begin by setting up some notation.

- Let P be the transition matrix of a time-reversible Markov chain on state space Ω with stationary distribution π ,

where $P[i, j]$ denotes the transition probability from i to j . Let $P_t = \exp(-t(I - P))$ denote the time- t transition kernel.

- For a distribution ν absolutely continuous with respect to π , we use $f(x) := \frac{d\nu}{d\pi}(x)$ to refer to its relative density to π .
- We use ν_t to denote $P_t\nu$, and we assume that ν_t is absolutely continuous with respect to π throughout. In particular, we write $f_t(x) := \frac{d\nu_t}{d\pi}(x)$ to denote its relative density to π ;
- We use $\mathbf{m}(\nu) := \mathbf{E}_{\mathbf{x} \sim \nu} \mathbf{x}$ to denote the mean of ν .
- We use $\mathbf{y} \sim_P \mathbf{x}$ when \mathbf{y} is chosen as a random neighbor of \mathbf{x} according to transition probabilities given by P . We drop the subscript P from the \sim when the Markov chain is clear from context.

Remark II.1. The way to think of the time- t transition kernel P_t for a Markov chain with kernel P on a discrete space is via the process: sample $t \sim \text{Poi}(t)$ and take t steps using P .

We will require the following simple consequence of the definition of total variation distance.

Fact II.2. For any pair of distributions ν and π on Ω , and any function $f : \Omega \rightarrow \mathbb{R}$, we have

$$|\mathbf{E}_\nu f - \mathbf{E}_\pi f| \leq (f_{\max} - f_{\min}) \cdot d_{\text{TV}}(\nu, \pi).$$

Definition II.3 (Dirichlet form). For functions $f, g : \Omega \rightarrow \mathbb{R}$, and $x, y \in \Omega$, the *Dirichlet form* of f and g with respect to P is:

$$\mathcal{E}_P(f, g) := \mathbf{E}_{\mathbf{x} \sim \pi} \mathbf{E}_{\mathbf{y} \sim_P \mathbf{x}} (f(\mathbf{x}) - f(\mathbf{y})) \cdot (g(\mathbf{x}) - g(\mathbf{y})).$$

We drop the P in the subscript when it is clear from context.

Remark II.4. When we use the Glauber dynamics chain for a distribution π on a hypercube, we use \mathcal{E}_π to denote the corresponding Dirichlet form.

The Dirichlet form measures the rate at which a Markov chain makes progress towards the stationary distribution. The following is one way of articulating this notion; see, e.g., [18].

Fact II.5. $\frac{d}{dt} \text{KL}(\nu_t \parallel \pi) = -\mathcal{E}(f_t, \log f_t) = -\mathbf{E}_{\mathbf{x} \sim \pi} \mathbf{E}_{\mathbf{y} \sim \mathbf{x}} (f_t(\mathbf{x}) - f_t(\mathbf{y})) \cdot \log \frac{f_t(\mathbf{x})}{f_t(\mathbf{y})}.$

Definition II.6 (Modified log-Sobolev inequality). We say P satisfies a *modified log-Sobolev inequality* (MLSI) with constant C if for any function $f : \Omega \rightarrow \mathbb{R}_{>0}$,

$$\mathcal{E}_P(f, \log f) \geq C \cdot \text{Ent}[f].$$

Here, $\text{Ent}[f] := \mathbf{E}_\pi[f \log f] - \mathbf{E}_\pi f \log \mathbf{E}_\pi f$ is the entropy functional. In particular, C_{MLSI} is the best (largest) such constant C .

We will need the following fact concerning the MLSI for product measures.

Fact II.7 (see e.g. [75, Lemma 2.5]). Let π be a distribution over $\{\pm 1\}^n$ with independent coordinates. Then $C_{\text{MLSI}}(\pi) \geq 1/n$.

Remark II.8. It is well known that the KL divergence to the stationary distribution decays exponentially at a rate dictated by $C_{\text{MLS}I}$ (see [18, Theorem 2.4] for more details):

$$\text{KL}(\nu_t\|\pi) \leq \text{KL}(\nu_0\|\pi) \cdot \exp(-C_{\text{MLS}I}t).$$

Measure decompositions. Some of the key properties of locally stationary distributions rely on the notion of a measure decomposition. These can be defined in great generality, but we will restrict our attention to distributions on subsets of \mathbb{R}^n for concreteness.

Definition II.9 (Measure decomposition). Let π be a distribution on \mathbb{R}^n . Let ρ be a mixture distribution, also on \mathbb{R}^n , which indexes into a family of mixture components $\{\pi_z\}_{z \in \mathbb{R}^n}$. We say that (ρ, π_z) is a measure decomposition for π if

$$\pi = \mathbf{E}_{z \sim \rho} \pi_z.$$

One should think of the mixture components π_z as being “simpler” distributions than the original measure π . Not all measure decompositions are useful; there is always a trivial measure decomposition where the mixture ρ is exactly π and the simpler distributions π_z are just Dirac masses at z .

Associated to each measure decomposition is a natural Markov chain; see, e.g., [48, Definition 6].

Definition II.10 (Markov chain associated to a measure decomposition). Given a measure decomposition $\pi = \mathbf{E}_{z \sim \rho} \pi_z$, its associated Markov chain is defined by

$$x \rightarrow z | x \rightarrow x' | z.$$

Notably, Glauber dynamics and restricted Gaussian dynamics can both be viewed as the associated Markov chain to certain measure decompositions. The relevant decomposition for Glauber dynamics represents π as the mixture of its conditional marginals.

Remark II.11. Measure decompositions constructed using stochastic localization have recently been used to prove functional inequalities for a wide class of Ising models [48], [50], [52].

Symmetric KL divergence. A useful potential function for us is the symmetric KL divergence.

Definition II.12. For a pair of distributions π and ν on Ω , we define their *symmetric KL divergence* as:

$$\text{SKL}(\pi, \nu) := \text{KL}(\nu\|\pi) + \text{KL}(\pi\|\nu).$$

Observation II.13. For any π and ν , setting f to be the density of ν with respect to π , we have

$$\text{SKL}(\pi, \nu) = \frac{1}{2} \cdot \mathbf{E}_{\mathbf{x}, \mathbf{y} \sim \pi} \left[(f(\mathbf{x}) - f(\mathbf{y})) \cdot \log \frac{f(\mathbf{x})}{f(\mathbf{y})} \right].$$

Observe that the above quantity is the Dirichlet form for the trivial “one-step” Markov chain with transition matrix $\mathbf{1}\pi^\top$ with stationary distribution π , which we shall denote $K(\pi)$.

We will also require the following inequality between the KL divergence and symmetric KL divergence, which we prove in the full version of the paper.

Lemma II.14. Let ν be an arbitrary distribution with density f with respect to π , and τ such that $\tau > \max_{x \in \Omega} \log f(x)$ or $\tau > \max_{x \in \Omega} \log \frac{1}{f(x)}$. Then, the symmetric KL divergence can be bounded in terms of the KL divergence as follows.

$$\text{SKL}(\pi, \nu) \leq (6 + 12\tau) \cdot \text{KL}(\nu\|\pi).$$

III. PROPERTIES OF LOCALLY STATIONARY DISTRIBUTIONS

We record some useful properties of locally stationary distributions below.

Random walks yield locally-stationary measures at a typical time. The following is a generic statement about any time-reversible Markov chain achieving a locally stationary distribution.

Lemma III.1. For any distribution ν , any Markov chain transition kernel P with stationary distribution π and any $T > 0$, for t chosen uniformly at random in $[0, T]$:

$$\mathbf{E}_{t \sim [0, T]} \mathcal{E}_P(f_t, \log f_t) \leq \frac{\text{KL}(\nu\|\pi)}{T} \leq \frac{\log \frac{1}{\pi_{\min}}}{T}. \quad (1)$$

Proof. By Fact II.5,

$$\begin{aligned} 0 &\leq \text{KL}(\nu_T\|\pi) \\ &= \text{KL}(\nu\|\pi) - \int_0^T \mathcal{E}(f_t, \log f_t) dt \\ &= \text{KL}(\nu\|\pi) - T \cdot \mathbf{E}_{t \sim [0, T]} \mathcal{E}(f_t, \log f_t) \\ &\leq \log \frac{1}{\pi_{\min}} - T \cdot \mathbf{E}_{t \sim [0, T]} \mathcal{E}(f_t, \log f_t). \end{aligned}$$

Rearranging the above gives us the desired statement. \square

A simple consequence of Lemma III.1 and Markov’s inequality is that for most times in $[0, T]$, ν_t is indeed locally stationary.

Theorem I.3. Fix a time-reversible Markov chain P with a stationary measure π , any starting distribution ν_0 , and $\varepsilon, \delta > 0$. Let $T = \frac{1}{\delta\varepsilon} \cdot \log \left(\frac{1}{\pi_{\min}} \right)$. Then, for a time $t \sim [0, T]$ chosen uniformly at random, the distribution ν_t at time t is ε -locally stationary with respect to P with probability at least $1 - \delta$.

a) *Stationarity over small time-scales.*: We will also require the observation that if the Dirichlet form at a distribution is small, so too is the total variation distance between it and the distribution obtained after one step of the Markov chain.

Lemma III.2. Let P be a reversible Markov chain with stationary distribution π , ν an arbitrary distribution, and f its density relative to π . Then,

$$\mathcal{E}_P(f, \log f) \geq 2 \cdot \text{KL}(P\nu\|\nu) \geq 4 \cdot d_{\text{TV}}(\nu, P\nu)^2.$$

Proof. We recall the definition of the Dirichlet form,

$$\mathcal{E}_P(f, \log f) = \mathbf{E}_{\mathbf{x} \sim \pi} \left[\mathbf{E}_{\mathbf{y} \sim P\mathbf{x}} (f(\mathbf{x}) - f(\mathbf{y})) \log \frac{f(\mathbf{x})}{f(\mathbf{y})} \right]$$

$$\begin{aligned}
&= 2 \cdot \mathbf{E}_{\substack{\mathbf{x} \sim \pi \\ \mathbf{y} \sim P\mathbf{x}}} f(\mathbf{x}) \log \left(\frac{f(\mathbf{x})}{f(\mathbf{y})} \right) \\
&= 2 \cdot \mathbf{E}_{\substack{\mathbf{x} \sim \nu \\ \mathbf{y} \sim P\mathbf{x}}} \log \left(\frac{f(\mathbf{x})}{f(\mathbf{y})} \right) \\
&= 2 \cdot (\mathbf{E}_{\mathbf{x} \sim \nu} \log f(\mathbf{x}) - \mathbf{E}_{\mathbf{y} \sim P\nu} \log f(\mathbf{y})),
\end{aligned}$$

where the second line used the reversibility of P . Adding and subtracting a term, we may neatly express the above in terms of KL divergences as

$$\begin{aligned}
\mathcal{E}_P(f, \log f) &= 2 \cdot \left[\left(\mathbf{E}_{\mathbf{x} \sim \nu} \log f(\mathbf{x}) - \mathbf{E}_{\mathbf{y} \sim P\nu} \log \frac{dP\nu}{d\pi}(\mathbf{y}) \right) \right. \\
&\quad \left. + \left(\mathbf{E}_{\mathbf{y} \sim P\nu} \log \frac{dP\nu}{d\pi}(\mathbf{y}) - \mathbf{E}_{\mathbf{y} \sim P\nu} \log f(\mathbf{y}) \right) \right] \\
&= 2 \cdot [(\text{KL}(\nu \| \pi) - \text{KL}(P\nu \| \pi)) + \text{KL}(P\nu \| \nu)] \\
&\geq 2 \cdot \text{KL}(P\nu \| \nu) \geq 4 \cdot d_{\text{TV}}(P\nu, \nu)^2
\end{aligned}$$

as desired, where the second-to-last inequality follows from the fact that the KL divergence to the stationary distribution is non-increasing with time, and the last inequality is Pinsker's. \square

Consequently, averages of bounded functions do not change much after one step of the Markov chain

Corollary III.3. *Let $\phi : \Omega \rightarrow \mathbb{R}$ be a bounded function on the state space of a Markov chain P , and ν be an ε -locally stationary measure. Then,*

$$|\mathbf{E}_{\mathbf{x} \sim \nu}[\phi(\mathbf{x})] - \mathbf{E}_{\mathbf{x} \sim P\nu}[\phi(\mathbf{x})]| \leq \|\phi\|_\infty \cdot \sqrt{\varepsilon}.$$

Locally stationary measures are close to stationary measures on small neighborhoods. As their name suggests, locally stationary distributions locally resemble the true stationary distribution. For example, typical samples from the locally stationary distribution approximately satisfies the detailed balance condition. Even though we do not explicitly employ this in any applications, we include it here as it gives the impression of a fundamental structural property of locally stationary distributions.

Lemma III.4. *For an ε -locally stationary distribution ν with relative density f , and for $\mathbf{x} \sim \nu$ and $\mathbf{y} \sim \mathbf{x}$, with probability at least $1 - \delta$, we have $\frac{f(\mathbf{x})}{f(\mathbf{y})} = 1 \pm O(\sqrt{\frac{\varepsilon}{\delta}})$, where $\delta > 2\varepsilon$.*

Proof. Since $\mathcal{E}(f, \log f) < \varepsilon$,

$$\begin{aligned}
\mathbf{E}_{\mathbf{x} \sim \pi} \mathbf{E}_{\mathbf{y} \sim \mathbf{x}} (f(\mathbf{x}) - f(\mathbf{y})) \cdot \log \frac{f(\mathbf{x})}{f(\mathbf{y})} &< \varepsilon \\
\mathbf{E}_{\mathbf{x} \sim \nu} \mathbf{E}_{\mathbf{y} \sim \mathbf{x}} \left(1 - \frac{f(\mathbf{y})}{f(\mathbf{x})} \right) \cdot \log \frac{f(\mathbf{x})}{f(\mathbf{y})} &< \varepsilon
\end{aligned}$$

Since the random variable at hand is always nonnegative, we can apply Markov's inequality, which tells us that with probability at least $1 - \delta$:

$$\left(1 - \frac{f(\mathbf{y})}{f(\mathbf{x})} \right) \cdot \log \frac{f(\mathbf{x})}{f(\mathbf{y})} < \frac{\varepsilon}{\delta}.$$

The claim then follows since the above inequality is violated if $\frac{f(\mathbf{x})}{f(\mathbf{y})}$ deviates from 1 by more than a constant multiple of $\sqrt{\frac{\varepsilon}{\delta}}$. \square

Formally, the following can be abstracted out of the proof of Theorem I.4:

Lemma III.5. *Let P be a Glauber dynamics chain for a distribution μ on $\{\pm 1\}^n$, and let ν be an ε -locally stationary measure with respect to P for some $\varepsilon > 0$. For a subset of coordinates $W \subset [n]$, and an assignment $x_{\bar{W}}$ of coordinates outside W , let $P_{W, x_{\bar{W}}}$ denote the Glauber dynamics chain of $\mu|_{x_{\bar{W}}}$. Suppose for every choice of W and $x_{\bar{W}}$, we have $C_{\text{MLSI}}(P_{W, x_{\bar{W}}}) \geq C$, then:*

$$\mathbf{E}_{\mathbf{x}_{\bar{W}} \sim \nu} [d_{\text{TV}}(\nu|_{\mathbf{x}_{\bar{W}}}, \pi|_{\mathbf{x}_{\bar{W}}})] \leq \frac{1}{C} \cdot \sqrt{\varepsilon}.$$

Corollary III.6. *In the setting of Lemma III.5, suppose $\phi : \{0, 1\}^W \rightarrow \mathbb{R}$ is a bounded function of x_W then,*

$$\mathbf{E}_{\mathbf{x} \sim \nu} [\phi(x_W)] \geq \mathbf{E}_{\mathbf{x}_{\bar{W}}} \mathbf{E}_{\mathbf{x}_W \sim \pi|_{x_{\bar{W}}}} [\phi(x_W)] - \frac{1}{C} \cdot \sqrt{\varepsilon} \cdot \|\phi\|_\infty.$$

Local stationarity is preserved over short time-scales of the random walk. The following lemma essentially says that if a distribution ν is locally stationary, then so is $P^T\nu$ for any small T .

Lemma III.7. *Let P be a Markov chain with stationary distribution π , and ν an arbitrary distribution with relative density f . Then,*

$$\mathcal{E}_{P^T}(f, \log f) \leq O(T^3) \cdot \mathcal{E}_P(f, \log f).$$

Proof. Suppose we pick $\mathbf{x} \sim \pi$, $\mathbf{y}_1 \sim_P \mathbf{x}$, then $\mathbf{y}_2 \sim_P \mathbf{y}_1$. Then, because π is stationary with respect to P ,

$$\begin{aligned}
\mathcal{E}_P(f, \log f) &= \mathbf{E} \left[(f(\mathbf{x}) - f(\mathbf{y}_1)) \log \frac{f(\mathbf{x})}{f(\mathbf{y}_1)} \right] \\
&= \mathbf{E} \left[(f(\mathbf{y}_1) - f(\mathbf{y}_2)) \log \frac{f(\mathbf{y}_1)}{f(\mathbf{y}_2)} \right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
2 \cdot \mathcal{E}_P(f, \log f) &= \mathbf{E} \left[(f(\mathbf{x}) - f(\mathbf{y}_1)) \log \frac{f(\mathbf{x})}{f(\mathbf{y}_1)} \right. \\
&\quad \left. + (f(\mathbf{y}_1) - f(\mathbf{y}_2)) \log \frac{f(\mathbf{y}_1)}{f(\mathbf{y}_2)} \right] \\
&\geq \frac{1}{4} \cdot \mathbf{E} \left[(f(\mathbf{x}) - f(\mathbf{y}_2)) \log \frac{f(\mathbf{x})}{f(\mathbf{y}_2)} \right] \\
&= \frac{1}{4} \cdot \mathcal{E}_{P^2}(f, \log f).
\end{aligned}$$

Here, the inequality follows from the observation that defining $d(p, q) = (p - q) \log(p/q)$, $d(p, q) + d(q, r) \geq \frac{1}{4} \cdot d(p, r)$. Indeed, if $q \notin (p, r)$, this is trivial, so assuming without loss of generality that $r \geq q \geq \frac{p+r}{2} \geq p$, we get that

$$d(p, q) \geq \left(\frac{p+r}{2} - p \right) \log \left(\frac{p+r}{2p} \right)$$

$$\begin{aligned}
&= \frac{1}{2} \cdot (r-p) \cdot \log \left(\frac{1}{2} \cdot \left(1 + \frac{r}{p} \right) \right) \\
&\geq \frac{1}{4} \cdot (r-p) \log \left(\frac{r}{p} \right).
\end{aligned}$$

Applying this $\log T$ times, it follows that

$$\mathcal{E}_{P^T}(f, \log f) \leq O(T^3) \cdot \mathcal{E}_P(f, \log f). \quad \square$$

Local stationarity can be transferred under component MLSI.

Let $\pi = \mathbf{E}_{z \sim \rho} \pi_z$ be a measure decomposition. In this section we prove that if Glauber dynamics is locally stationary, so too is the Markov chain associated with the measure decomposition, provided that the mixture components π_z all have a good MLSI constant.

Lemma III.8. *Let P be the Markov chain associated to a measure decomposition $\pi = \mathbf{E}_{z \sim \rho} \pi_z$. Let $f : \{\pm 1\}^n \rightarrow \mathbb{R}_{>0}$ be any function and set τ such that $\min_{x \in \{\pm 1\}^n} f(x) > \exp(-\tau)$ or $\max_{x \in \{\pm 1\}^n} f(x) < \exp(\tau)$. For $\delta := \inf_z C_{\text{MLSI}}(\pi_z)$, we have*

$$\mathcal{E}_P(f, \log f) \leq O\left(\frac{\tau}{\delta}\right) \cdot \mathcal{E}_\pi(f, \log f).$$

Proof. We use \mathcal{C}_n to denote the hypercube graph on vertex set $\{\pm 1\}^n$ with edge set having pairs of vertices that differ in a single coordinate. For any nonnegative function f and distribution π with $\mathbf{E}_\pi f = 1$, we use $f \cdot \pi$ to denote the distribution ν with $\frac{d\nu}{d\pi}(x) = f(x)$.

For any function f satisfying the assumption of the statement, let $\mathcal{E}_{xy}(f, \log f) := (f(x) - f(y)) \log \frac{f(x)}{f(y)}$. Then we have

$$\begin{aligned}
&\mathcal{E}_\pi(f, \log f) \\
&= \sum_{\{x,y\} \in \mathcal{C}_n} \frac{1}{n} \cdot \frac{\pi(x) \cdot \pi(y)}{\pi(x) + \pi(y)} \cdot \mathcal{E}_{xy}(f, \log f) \\
&= \sum_{\{x,y\} \in \mathcal{C}_n} \frac{1}{n} \cdot \frac{\mathbf{E}_{z \sim \rho} \pi_z(x) \cdot \mathbf{E}_{z \sim \rho} \pi_z(y)}{\mathbf{E}_{z \sim \rho} \pi_z(x) + \mathbf{E}_{z \sim \rho} \pi_z(y)} \cdot \mathcal{E}_{xy}(f, \log f) \\
&\geq \sum_{\{x,y\} \in \mathcal{C}_n} \frac{1}{n} \cdot \mathbf{E}_{z \sim \rho} \left[\frac{\pi_z(x) \cdot \pi_z(y)}{\pi_z(x) + \pi_z(y)} \right] \cdot \mathcal{E}_{xy}(f, \log f) \\
&= \mathbf{E}_{z \sim \rho} [\mathcal{E}_{\pi_z}(f, \log f)].
\end{aligned}$$

Above, the inequality follows from the concavity of the function $(a, b) \mapsto \frac{ab}{a+b}$ in the non-negative quadrant, and all the Dirichlet forms are with respect to the Glauber dynamics chain.

Continuing the above calculation,

$$\begin{aligned}
&\mathbf{E}_{z \sim \rho} [\mathcal{E}_{\pi_z}(f, \log f)] \\
&\geq \mathbf{E}_{z \sim \rho} \left[C_{\text{MLSI}}(\pi_z) \cdot \mathbf{E}_{\pi_z}[f] \cdot \text{KL} \left(\frac{f}{\mathbf{E}_{\pi_z} f} \cdot \pi_z \parallel \pi_z \right) \right] \\
&\geq \Omega\left(\frac{\delta}{\tau}\right) \mathbf{E}_{z \sim \rho} \left[\mathbf{E}_{\pi_z}[f] \cdot \text{SKL} \left(\frac{f}{\mathbf{E}_{\pi_z} f} \cdot \pi_z, \pi_z \right) \right] \\
&= \Omega\left(\frac{\delta}{\tau}\right) \mathbf{E}_{z \sim \rho} \left[\mathbf{E}_{x,y \sim \pi_z} (f(x) - f(y)) \log \frac{f(x)}{f(y)} \right] \\
&= \Omega\left(\frac{\delta}{\tau}\right) \mathcal{E}_P(f, \log f),
\end{aligned}$$

where the second inequality follows by Lemma II.14 and MLSI, and the third line follows from Observation II.13. The claim follows. \square

The upshot is that we have complete freedom to select the measure decomposition, provided we can establish an MLSI for the components. This can be useful when it is easier to directly analyze the consequences of local stationarity for the associated Markov chain instead of Glauber dynamics.

IV. WARMUP: LARGE INDEPENDENT SETS IN TRIANGLE-FREE GRAPHS

Observe that any graph G on n vertices with maximum degree d has an independent set of size $\frac{n}{d+1}$, a bound which is tight for the disjoint union of $(d+1)$ -sized cliques. Ajtai, Komlós, and Szemerédi [19] showed that when G is triangle-free, the size of the maximum independent set increases to $\Omega(n \cdot \frac{\log d}{d})$. Shearer [20] gave an alternate proof that shows such an independent set exists with a leading constant of 1, even if G merely has average-degree bounded by d . As a warmup, we prove that Glauber dynamics succeeds at finding a large independent set in $O(nd^4)$ steps.

Theorem I.4. *Let G be a triangle-free graph on n vertices with maximum degree bounded by d . Let \mathcal{I} be an independent set in G that arises from Glauber dynamics run for $O(nd^4)$ time. Then the expected size of \mathcal{I} is at least $\frac{1-o_d(1)}{4} \cdot n \cdot \frac{\log d}{d}$.*

To prove Theorem I.4, we will need the following crude bound on the modified log-Sobolev constant for the uniform distribution over independent sets of a star. A short proof is provided at the end of this section.

Lemma IV.1. *Let π denote the uniform distribution over independent sets of a star with Δ many leaves. Then $C_{\text{MLSI}}(\pi) \geq \exp(-O(\Delta))$.*

Remark IV.2. The bound can easily be made $\frac{1}{\text{poly}(\Delta)}$, but we will not need this here.

We also leverage the following simple and well-known lemma on the local behavior of a uniformly random independent set. For completeness, we include a short proof of it at the end of this section, following the one provided in Alon & Spencer [21, Proposition 1, Page 272]. Throughout this section, we write $N(v) = \{u \in V : u \sim v\}$ for the open neighborhood of $v \in V$, and $N[v] = N(v) \cup \{v\}$ for the closed neighborhood.

Lemma IV.3. *Let G be a triangle-free graph of maximum degree d , and let π denote the uniform measure over independent sets of G . For every vertex $v \in V$, define the following real-valued score function over $\{0, 1\}^V$:*

$$\phi_v(\mathbf{x}) := d\mathbf{x}_v + \sum_{u \in N(v)} \mathbf{x}_u. \quad (2)$$

Then for every pinning $\tau \in \{0, 1\}^{\overline{N[v]}}$, we have

$$\mathbf{E}_{\mathbf{x} \sim \pi} [\phi_v(\mathbf{x}) \mid \mathbf{x}_{\overline{N[v]}} = \tau] \geq \frac{\log d}{2}.$$

The key property of this score function is that it readily yields a lower bound on the size of an independent set $x \in \{0, 1\}^V$. This follows from the observation that

$$n \cdot \mathbf{E}_{v \sim V} \phi_v(\mathbf{x}) \leq 2d \cdot \sum_{v \in V} \mathbf{x}_v. \quad (3)$$

Note that by averaging over $\tau \in \{0, 1\}^{\overline{N[v]}}$ drawn from the marginal distribution of π induced on $\overline{N[v]}$, the conclusion of Lemma IV.3 combined with Eq. (3) implies that a uniformly random independent set drawn from π has expected size at least $\frac{1}{4} \cdot n \cdot \frac{\log d}{d}$. We observe that the same claim holds even if the distribution over independent sets is merely locally stationary with respect to Glauber dynamics, rather than being truly uniform.

Proof of Theorem I.4. As discussed above, a direct application of the law of total expectation combined with Lemma IV.3 yields the first claim concerning the expected size of a uniformly random independent set. We now turn to the second claim. Let $T \geq 0$ be a parameter to be determined later, and for every $0 \leq t \leq T$, let ν_t denote the distribution over independent sets after running Glauber dynamics for time- t from an arbitrary initialization. Our goal is to establish the lower bound

$$\mathbf{E}_{t \sim [0, T]} \mathbf{E}_{\nu_t} \mathbf{E}_{v \sim V} \phi_v(\mathbf{x}) \geq \frac{\log d}{2} - \varepsilon \quad (4)$$

for $0 < \varepsilon < o_d(1)$, which when combined with Eq. (3) immediately implies that the expected size of the independent set discovered by Glauber dynamics is $\left(\frac{1-o_d(1)}{4}\right) \cdot n \cdot \frac{\log d}{d}$. For the purpose of analysis, if ν is any distribution over independent sets, we shall think of $\mathbf{x} \sim \nu$ as being sampled in the following alternate way.

- 1) For a fixed vertex v , sample $\mathbf{x}_{\overline{N[v]}}$ from the marginal distribution induced by ν on $\overline{N[v]}$. For each $w \in N(v)$ that has a neighbor in the independent set $\mathbf{x}_{\overline{N[v]}}$, pin \mathbf{x}_w to 0, since it is deterministically equal to 0 in the conditional measure $\nu|\mathbf{x}_{\overline{N[v]}}$.
- 2) If the number of unpinned vertices at this stage is strictly larger than $\log d$, sample \mathbf{x}_v from its corresponding conditional marginal distribution.
- 3) Let U be the set of remaining unpinned vertices. Sample $\mathbf{x}_U \sim \nu|\mathbf{x}_{\overline{U}}$.

For any vertex $v \in V$, we have

$$\begin{aligned} \mathbf{E}_{\mathbf{x} \sim \nu} \phi_v(\mathbf{x}) &= \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} \mathbf{E}_{\nu \mid \mathbf{x}_{\overline{U}}} \phi_v(\mathbf{x}) \\ &\geq \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} [\mathbf{E}_{\pi \mid \mathbf{x}_{\overline{U}}} \phi_v(\mathbf{x}) - 2d \cdot d_{\text{TV}}(\nu \mid \mathbf{x}_{\overline{U}}, \pi \mid \mathbf{x}_{\overline{U}})] \\ &\geq \frac{\log d}{2} - 2d \cdot \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} d_{\text{TV}}(\nu \mid \mathbf{x}_{\overline{U}}, \pi \mid \mathbf{x}_{\overline{U}}), \end{aligned} \quad (5)$$

Note that the random subset of vertices U , as well as the boundary condition $\mathbf{x}_{\overline{U}}$, are all drawn from the above process with respect to ν , not π . The first inequality follows by applying Fact II.2 along with $2d$ -boundedness and nonnegativity of the score function ϕ_v . For the second inequality, note that if $v \in U$, then we may invoke Lemma IV.3. Now suppose $v \notin U$. If v is pinned 1, then $\phi_v(\mathbf{x}) = d$. If v is pinned 0, then by

triangle-freeness, $\mathbf{x}_u = 1$ with probability 1/2 independently for all $u \in U$. Since $|U| \geq \log d$, the lower bound follows.

In the rest of this argument, we will show that when $t \sim [0, T]$, ν is equal to ν_t and v is chosen uniformly at random, we can achieve a strong upper bound on

$$\mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} d_{\text{TV}}(\nu_t \mid \mathbf{x}_{\overline{U}}, \pi \mid \mathbf{x}_{\overline{U}}).$$

For the rest of this proof, we shall abbreviate $\nu_t \mid \mathbf{x}_{\overline{U}}$ and $\pi \mid \mathbf{x}_{\overline{U}}$ as ν'_t and π' , respectively. Furthermore, let f'_t denote the relative density of ν'_t with respect to π' . By Pinsker's inequality, we can bound the above by:

$$\mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} \sqrt{\text{KL}(\nu'_t \mid\mid \pi')} \leq \sqrt{\mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} \text{KL}(\nu'_t \mid\mid \pi')}.$$

We focus our attention on showing an upper bound on the term in the square-root.

$$\begin{aligned} &\mathbf{E}_t \mathbf{E}_{v \sim V} \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} \text{KL}(\nu'_t \mid\mid \pi') \\ &\leq \mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} \frac{1}{C_{\text{MLSI}}(\pi')} \mathcal{E}_{\pi'}(f'_t, \log f'_t) \\ &\leq O(d) \cdot \mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, \mathbf{x}_{\overline{U}} \mid v} \mathcal{E}_{\pi'}(f'_t, \log f'_t) \\ &\leq O(d^2) \cdot \mathbf{E}_t \mathcal{E}_{\pi}(f_t, \log f_t) \\ &\leq O(d^2) \cdot \frac{n}{T}. \end{aligned}$$

In the above, the first inequality uses the definition of $C_{\text{MLSI}}(\pi')$. For the second inequality, note that almost surely, either $G[U]$ is a star centered at v with at most $\log d$ many leaves, or $G[U]$ consists entirely of isolated vertices due to pinning v . In either case, we have $C_{\text{MLSI}}(\pi') \geq \Omega(\frac{1}{d})$ by appropriately applying Fact II.7 or Lemma IV.1. The third inequality is based on comparing Dirichlet forms. The final inequality is a direct application of Lemma III.1. Plugging in the above into Eq. (5) and setting $T = O(nd^4/\varepsilon^2)$, we get Eq. (4) as desired. \square

Remark IV.4. Using a similar argument, one can establish a similar result for Max-Cut on triangle-free graphs with maximum degree d . In particular, Glauber dynamics run for $\text{poly}(n)$ many steps on the antiferromagnetic Ising model on G with inverse temperature $\frac{1}{\sqrt{d}}$ finds a cut of size $\frac{1}{2} + \Omega\left(\frac{1}{\sqrt{d}}\right)$.

Remark IV.5. Similar to the proof of weak recovery, a proof of the above can also be recast from the perspective of transferring local stationarity from Glauber dynamics to a new Markov chain that is more amenable to analysis (Lemma III.8). A single step of this new Markov chain picks a random vertex v , pins all spins except the ones on v and its neighbors, and then rerandomizes the spins on v and its neighbors according to the conditional distribution. The associated measure decomposition is that which picks a random vertex and conditions on everything except the vertex and its neighbors — as desired, the component measures indeed satisfy polynomial MLSIs. The “local proof” that the stationary distribution attains large objective value then immediately implies that one step of this Markov chain significantly increases the expected size of an independent set, if the expected size was much smaller than

$\Omega\left(\frac{\log d}{d} \cdot n\right)$. Corollary III.3 implies that this is not possible for a locally stationary distribution, a contradiction.

However, we do not present the proof in this form because as stated, it would lose an additional factor of n in the running time of the algorithm.¹

Proof of Lemma IV.3. Let $S \subseteq N(v)$ denote the subset of neighbors of v which are not adjacent to any vertex of the independent set $\mathbf{x}_{\overline{N[v]}} = \tau$, and write $k = |S|$. Observe that the distribution of $\mathbf{x}_{N[v]}$ conditioned on $\mathbf{x}_{\overline{N[v]}} = \tau$ is given by choosing the singleton $\{v\}$ with probability $\frac{1}{2^k+1}$, or a uniformly random subset of S with the remaining probability. Hence,

$$\mathbf{E}_{\mathbf{x} \sim \pi}[\phi_v(\mathbf{x}) \mid \mathbf{x}_{\overline{N[v]}} = \tau] = \frac{d}{2^k+1} + \frac{k}{2} \cdot \frac{2^k}{2^k+1}.$$

The above expression is always at least $\frac{\log d}{2}$ for any choice of nonnegative integer k . \square

Proof of Lemma IV.1. By [76, Corollary A.4] and [18, Proposition 3.6], we have that $C_{\text{MLSI}}(\pi) \geq \frac{1-2\pi_*}{\log(\frac{1}{\pi_*}-1)} \cdot \lambda(\pi)$, where $\lambda(\pi)$ denotes the spectral gap of Glauber dynamics for π , and $\pi_* = \min_{\mathbf{x}: \pi(\mathbf{x}) > 0} \pi(\mathbf{x}) = \frac{1}{2^{\Delta+1}}$. Hence, it suffices to establish that $\lambda(\pi) \geq \exp(-O(\Delta))$. For this, we appeal to the simple fact that random walk on a connected graph with n vertices has spectral gap at least $1/\text{poly}(n)$. A comparison of Dirichlet forms between Glauber dynamics and simple random walk on the $n = 2^\Delta + 1$ many independent sets of G yields the desired lower bound. \square

REFERENCES

- [1] R. Montenegro and P. Tetali, “Mathematical Aspects of Mixing Times in Markov Chains,” *Foundations and Trends in Theoretical Computer Science*, vol. 1, no. 3, pp. 237–354, 2006.
- [2] P. Diaconis, “The Markov Chain Monte Carlo Revolution,” *Bull. Amer. Math. Soc.*, vol. 46, no. 2, 2009.
- [3] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*. CRC Press LLC, 2011.
- [4] D. Bakry, I. Gentil, M. Ledoux *et al.*, *Analysis and geometry of Markov diffusion operators*. Springer, 2014, vol. 103.
- [5] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, 2nd ed. American Mathematical Society, 2017.
- [6] A. Sly, “Computational Transition at the Uniqueness Threshold,” in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 2010, pp. 287–296.
- [7] A. Sly and N. Sun, “The Computational Hardness of Counting in Two-Spin Models on d -Regular Graphs,” *The Annals of Probability*, vol. 42, no. 6, pp. 2383–2416, 2014.
- [8] O. Guédon and R. Vershynin, “Community detection in sparse networks via Grothendieck’s inequality,” *Probability Theory and Related Fields*, vol. 165, no. 3–4, pp. 1025–1049, 2016.
- [9] B. Selman, H. A. Kautz, and B. Cohen, “Noise strategies for improving local search,” in *AAAI*, vol. 94, 1994, pp. 337–343.
- [10] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, “Local entropy as a measure for sampling solutions in constraint satisfaction problems,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2016, no. 2, p. 023301, 2016.
- [11] M. Mørup and M. N. Schmidt, “Bayesian community detection,” *Neural computation*, vol. 24, no. 9, pp. 2434–2456, 2012.
- [12] J. Geng, A. Bhattacharya, and D. Pati, “Probabilistic community detection with unknown number of communities,” *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 893–905, 2019.
- [13] Y. Nesterov, *Lectures on convex optimization*. Springer, 2018, vol. 137.
- [14] R. Ge, C. Jin, and Y. Zheng, “No spurious local minima in nonconvex low rank problems: A unified geometric analysis,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1233–1242.
- [15] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, “Benign overfitting in linear regression,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.
- [16] Z. Ji and M. Telgarsky, “Directional convergence and alignment in deep learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 176–17 186, 2020.
- [17] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, “On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points,” *Journal of the ACM (JACM)*, vol. 68, no. 2, pp. 1–29, 2021.
- [18] S. G. Bobkov and P. Tetali, “Modified logarithmic Sobolev inequalities in discrete settings,” *Journal of Theoretical Probability*, vol. 19, no. 2, pp. 289–336, 2006.
- [19] M. Ajtai, J. Komlós, and E. Szemerédi, “A note on Ramsey numbers,” *Journal of Combinatorial Theory, Series A*, vol. 29, no. 3, pp. 354–360, 1980.
- [20] J. B. Shearer, “A note on the independence number of triangle-free graphs,” *Discrete Mathematics*, vol. 46, no. 1, pp. 83–87, 1983.
- [21] N. Alon and J. H. Spencer, *The Probabilistic Method*. John Wiley & Sons, 2016.
- [22] E. Mossel, D. Weitz, and N. C. Wormald, “On the hardness of sampling independent sets beyond the tree threshold,” *Probability Theory and Related Fields*, vol. 143, pp. 401–439, 2007.
- [23] A. Galanis, D. Štefankovič, and E. Vigoda, “Inapproximability for Antiferromagnetic Spin Systems in the Tree Nonuniqueness Region,” *J. ACM*, vol. 62, no. 6, dec 2015. [Online]. Available: <https://doi.org/10.1145/2785964>
- [24] E. Davies, M. Jansen, W. Perkins, and B. Roberts, “On the average size of independent sets in triangle-free graphs,” *Proceedings of the American Mathematical Society*, vol. 146, no. 1, pp. 111–124, 2018.
- [25] J. Baik, G. B. Arous, and S. Péché, “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices,” *The Annals of Probability*, vol. 33, no. 5, pp. 1643 – 1697, 2005. [Online]. Available: <https://doi.org/10.1214/009117905000000233>
- [26] M. Dia, N. Macris, F. Krzakala, T. Lesieur, L. Zdeborová *et al.*, “Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [27] ———, “Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [28] Y. Deshpande, E. Abbe, and A. Montanari, “Asymptotic mutual information for the balanced binary stochastic block model,” *Information and Inference: A Journal of the IMA*, vol. 6, no. 2, pp. 125–170, 2017.
- [29] A. El Alaoui and F. Krzakala, “Estimation in the spiked wigner model: a short proof of the replica formula,” in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1874–1878.
- [30] M. Lelarge and L. Miolane, “Fundamental limits of symmetric low-rank matrix estimation,” in *Conference on Learning Theory*. PMLR, 2017, pp. 1297–1301.
- [31] L. Miolane, “Fundamental limits of low-rank matrix estimation: the non-symmetric case,” *arXiv preprint arXiv:1702.00473*, 2017.
- [32] J. Barbier and N. Macris, “The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference,” *Probability theory and related fields*, vol. 174, pp. 1133–1185, 2019.
- [33] A. Montanari and R. Venkataraman, “Estimation of low-rank matrices via approximate message passing,” *The Annals of Statistics*, vol. 49, no. 1, pp. 321 – 345, 2021. [Online]. Available: <https://doi.org/10.1214/20-AOS1958>
- [34] Z. Fan, S. Mei, and A. Montanari, “TAP free energy, spin glasses and variational inference,” *The Annals of Probability*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:53374945>
- [35] M. Celentano, Z. Fan, and S. Mei, “Local convexity of the TAP free energy and AMP convergence for Z2-synchronization,” *The Annals of Statistics*, vol. 51, no. 2, pp. 519–546, 2023.
- [36] A. Montanari and Y. Wu, “Posterior sampling from the spiked models via diffusion processes,” *arXiv preprint arXiv:2304.11449*, 2023.
- [37] S. R. Allen, R. O’Donnell, and D. Witmer, “How to refute a random CSP,” in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE, 2015, pp. 689–708.

¹The additional factor of n comes from the extra factor of τ in Lemma III.8.

[38] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications,” *Physical review E*, vol. 84, no. 6, p. 066106, 2011.

[39] H. Kesten and B. P. Stigum, “Additional limit theorems for indecomposable multidimensional galton-watson processes,” *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1463–1481, 1966.

[40] ———, “Limit theorems for decomposable multi-dimensional galton-watson processes,” *Journal of Mathematical Analysis and Applications*, vol. 17, no. 2, pp. 309–338, 1967.

[41] E. Mossel, J. Neeman, and A. Sly, “A proof of the block model threshold conjecture,” *Combinatorica*, vol. 38, no. 3, p. 665–708, jun 2018. [Online]. Available: <https://doi.org/10.1007/s00493-016-3238-8>

[42] L. Massoulié, “Community detection thresholds and the weak Ramanujan property,” in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2014, pp. 694–703.

[43] E. Mossel, J. Neeman, and A. Sly, “Reconstruction and estimation in the planted partition model,” *Probability Theory and Related Fields*, vol. 162, no. 3-4, pp. 431–461, 2015.

[44] E. Abbe, “Community detection and stochastic block models: recent developments,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.

[45] C. Moore, “The computer science and physics of community detection: Landscapes, phase transitions, and hardness,” *arXiv preprint arXiv:1702.00467*, 2017.

[46] Y. T. Lee, R. Shen, and K. Tian, “Structured logconcave sampling with a restricted Gaussian oracle,” in *Conference on Learning Theory*. PMLR, 2021, pp. 2993–3050.

[47] R. Shen, K. Tian, and Y. T. Lee, “Composite Logconcave Sampling with a Restricted Gaussian Oracle,” *arXiv preprint arXiv:2006.05976*, 2020.

[48] Y. Chen and R. Eldan, “Localization Schemes: A Framework for Proving Mixing Bounds for Markov Chains (extended abstract),” in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. Los Alamitos, CA, USA: IEEE Computer Society, nov 2022, pp. 110–122. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/FOCS54457.2022.00018>

[49] R. Bauerschmidt and T. Bodineau, “A very simple proof of the LSI for high temperature spin systems,” *Journal of Functional Analysis*, vol. 276, no. 8, pp. 2582–2588, 2019.

[50] R. Eldan, F. Koehler, and O. Zeitouni, “A spectral condition for spectral gap: fast mixing in high-temperature Ising models,” *Probability theory and related fields*, vol. 182, no. 3, pp. 1035–1051, 2022.

[51] N. Anari, V. Jain, F. Koehler, H. T. Pham, and T.-D. Vuong, “Entropic independence I: Modified log-Sobolev inequalities for fractionally log-concave distributions and high-temperature ising models,” *arXiv preprint arXiv:2106.04105*, 2021.

[52] K. Liu, S. Mohanty, A. Rajaraman, and D. X. Wu, “Fast Mixing in Sparse Random Ising Models,” *arXiv preprint arXiv:2405.06616*, 2024.

[53] J. Hubbard, “Calculation of Partition Functions,” *Phys. Rev. Lett.*, vol. 3, pp. 77–78, Jul 1959. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.3.77>

[54] F. Koehler, H. Lee, and A. Risteski, “Sampling approximately low-rank Ising models: MCMC meets variational methods,” in *Conference on Learning Theory*. PMLR, 2022, pp. 4945–4988.

[55] A. Bovier and F. Den Hollander, *Metastability: a potential-theoretic approach*. Springer, 2016, vol. 351.

[56] K. Balasubramanian, S. Chewi, M. A. Erdogdu, A. Salim, and S. Zhang, “Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo,” in *Conference on Learning Theory*. PMLR, 2022, pp. 2896–2923.

[57] X. Cheng, B. Wang, J. Zhang, and Y. Zhu, “Fast Conditional Mixing of MCMC Algorithms for Non-log-concave Distributions,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[58] D. Kunisky, “Optimality of Glauber dynamics for general-purpose Ising model sampling and free energy approximation,” *arXiv preprint arXiv:2307.12581*, 2023.

[59] Z. Chen, E. Mossel, and I. Zadik, “Almost-Linear Planted Cliques Elude the Metropolis Process,” in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2023, pp. 4504–4539.

[60] R. Gheissari, A. Jagannath, and Y. Xu, “Finding planted cliques using Markov chain Monte Carlo,” *arXiv preprint arXiv:2311.07540*, 2023.

[61] M. Sellke, “The threshold energy of low temperature langevin dynamics for pure spherical spin glasses,” *Communications on Pure and Applied Mathematics*, 2023.

[62] S. B. Hopkins and D. Steurer, “Efficient Bayesian estimation from few samples: community detection and related problems,” in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2017, pp. 379–390.

[63] A. El Alaoui, A. Montanari, and M. Sellke, “Shattering in pure spherical spin glasses,” *arXiv preprint arXiv:2307.04659*, 2023.

[64] G. Parisi, “A sequence of approximated solutions to the SK model for spin glasses,” *Journal of Physics A: Mathematical and General*, vol. 13, no. 4, p. L115, 1980.

[65] M. Talagrand, “The Parisi formula,” *Annals of mathematics*, pp. 221–263, 2006.

[66] M. Mézard, G. Parisi, and R. Zecchina, “Analytic and algorithmic solution of random satisfiability problems,” *Science*, vol. 297, no. 5582, pp. 812–815, 2002.

[67] S. Mertens, M. Mézard, and R. Zecchina, “Threshold values of random K-SAT from the cavity method,” *Random Structures & Algorithms*, vol. 28, no. 3, pp. 340–373, 2006.

[68] J. Ding, A. Sly, and N. Sun, “Proof of the satisfiability conjecture for large k ,” in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, 2015, pp. 59–68.

[69] A. Coja-Oghlan, F. Krzakala, W. Perkins, and L. Zdeborová, “Information-theoretic thresholds from the cavity method,” in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 2017, pp. 146–157.

[70] A. Coja-Oghlan and W. Perkins, “Bethe states of random factor graphs,” *Communications in Mathematical Physics*, vol. 366, pp. 173–201, 2019.

[71] ———, “Spin systems on Bethe lattices,” *Communications in Mathematical Physics*, vol. 372, pp. 441–523, 2019.

[72] R.-D. Buhai, P. K. Kothari, and D. Steurer, “Algorithms approaching the threshold for semi-random planted clique,” in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023, pp. 1918–1926.

[73] V. Guruswami, J.-T. Hsieh, P. K. Kothari, and P. Manohar, “Efficient Algorithms for Semirandom Planted CSPs at the Refutation Threshold,” in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2023, pp. 307–327.

[74] B. Barak, P. Raghavendra, and D. Steurer, “Rounding semidefinite programming hierarchies via global correlation,” in *2011 ieee 52nd annual symposium on foundations of computer science*. IEEE, 2011, pp. 472–481.

[75] S. Goel, “Modified logarithmic Sobolev inequalities for some models of random walk,” *Stochastic processes and their applications*, vol. 114, no. 1, pp. 51–79, 2004.

[76] P. Diaconis and L. Saloff-Coste, “Logarithmic Sobolev inequalities for finite Markov chains,” *The Annals of Applied Probability*, vol. 6, no. 3, 8 1996.